On the Relationship between Similarity Measures and Thresholds of Statistical Significance in the Context of Comparing Fuzzy Sets

Josie McCulloch *(Member, IEEE)*, Zack Ellerby *(Member, IEEE)* and Christian Wagner *(Senior Member, IEEE)*

***Abstract*—Comparing fuzzy sets by computing their similarity is common, with a large set of measures of similarity available. However, while commonplace in the computational intelligence community, the application and results of similarity measures are less common in the wider scientific context, where statistical approaches are the standard for comparing distributions. This is challenging, as it means that developments around similarity measures arising from the fuzzy community are inaccessible to the wider scientific community; and that the fuzzy community fails to take advantage of a strong statistical understanding which may be applicable to comparing (fuzzy membership) functions. In this paper, we commence a body of work on systematically relating the outputs of similarity measures to the notion of statistically significant difference; that is, how (dis)similar do two fuzzy sets need to be for them to be statistically different? We explain that in this context it is useful to initially focus on dis-similarity, rather than similarity, as the former aligns directly with the widely used concept of statistical difference. We propose two methods of applying statistical tests to the outputs of fuzzy dissimilarity measures to determine significant difference. We show how the proposed work provides deeper insight into the behaviour and possible interpretation of degrees of dis-similarity and, consequently, similarity, and how the interpretation differs in respect to context (e.g., the complexity of the fuzzy sets).**

I. Introduction

Similarity measures have been developed to compare fuzzy sets for a wide range of applications, including approximate reasoning [1, 2], clustering [3, 4], pattern recognition [5], image recognition [6], data mining [7], fuzzy rule-base simplification [8, 9], risk analysis [10, 11] and computing with words [12–14]. A large number of measures have been published with their specific properties tailored to specific applications' needs. As a consequence, different methods of measuring similarity provide different results when applied to the same sets. This is one factor that illustrates why it is often difficult to understand when and how the result of a given measure is meaningful; for example, a value x may be considered a meaningful degree of similarity for one measure but not for another.

Similarity measures have largely been developed in isolation, interacting little with the wider academic literature where statistical comparison is the norm. Intuitively, similarity measures can be helpful in problems outside the fuzzy sets and systems domains, helping to systematically compare distributions arising in areas from the social sciences to psychology. At the same time, it seems equally evident that the literature on similarity measures can directly benefit from the strong body of work on assessing whether data sets are statistically significantly different or not.

In the wider scientific literature, the statistical significance of a difference is the standard comparison between data sets. Without understanding the relationship between the outputs of similarity measures and whether or not a difference is found to be significant, we cannot relate the traditional statistical measures to measures of similarity, and cannot assess whether a given degree of similarity is objectively *meaningful*; that is, whether there is no statistically significant difference between them.

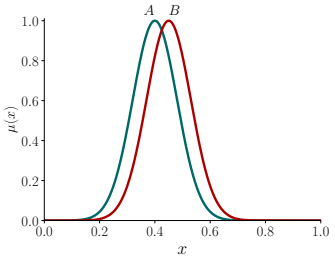
As an example, Fig. 1 shows two similarity measures, Jaccard s_J and Sørensen-Dice s_D (hereafter referred to as Dice), and their results when comparing the given fuzzy sets. Their results differ and it is difficult to assess whether the value $s_J = 0.6058$ represents meaningful similarity or a statistically significant difference. To illustrate - consider both sets to represent fuzzy set models of words (as in the Computing with Words paradigm): an obvious question is whether both words are the 'same' and how much so. Across the sciences, statistical approaches would be used to compare the underlying distributions of the models, while in the fuzzy set community, similarity measures would be employed.

We focus on finding a threshold σ, where, in the case of Jaccard, $s_J \geq \sigma_J$ would indicate meaningful similarity and $s_J < \sigma_J$ non-meaningful similarity. We also investigate the threshold for other similarity measures, which we expect to differ from σ_J. Fig. 2 shows an example of these thresholds for S_J and S_D.

There are many applications in which such thresholds are useful. For example, Navarro et al. [12] use a similarity measure to compare the meanings of words used by patients and medical professionals to measure the progress of patients' abilities. Without knowing a threshold of meaningful similarity, it is not clear if the patients and professionals share the same understanding of a word or if they have different interpretations. Similarity measures have also been applied to simplify fuzzy logic systems, to ensure an appropriate partitioning of the input or output space [9, 15]. Each fuzzy set is compared with each other, and if their similarity is higher than a given threshold then they are considered too similar, resulting in a system with redundant rules. In this case,

This work funded in part by the EPSRC's EP/M50810X/1 and in part by EP/P011918/1.

The authors are with LUCID - Lab for Uncertainty in Data and Decision Making, School of Computer Science, University of Nottingham, United Kingdom (email: josie.mcculloch; zack.ellerby; christian.wagner@nottingham.ac.uk



Measure	Result
$s_J(A,B) = \frac{A \cap B}{A \cup B}$	0.6058
$s_D(A,B) = \frac{2(A \cap B)}{\lvert A \rvert + \lvert B \rvert}$	0.7545

Fig. 1. Two fuzzy sets (with means 0.4 (A) and 0.45 (B), and equal std. dev. of 0.08) and their similarity according to the Jaccard (s_J) and Sørensen-Dice (s_D) measures.

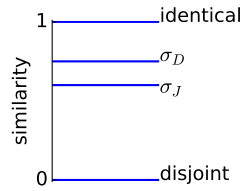


Fig. 2. Thresholds of meaningful results for similarity measures. The values 0 and 1 indicate disjoint and identical sets, respectively. A given value σ_J represents the threshold of meaningful results for the Jaccard similarity measure. The value σ_D – the threshold for Dice, may be different to σ_J.

it is essential to find an appropriate threshold of significant similarity, where fuzzy sets below this threshold are commonly either merged or one of them is removed.

In the context of ranking, the ordering of similarity values is conventionally more important than the values themselves [16]. Nevertheless, the actual values of similarity are useful if a threshold is known. For example, in classification we may assign a fuzzy set to the class with which it has the highest similarity - often regardless of the actual value of similarity. However, it is useful to understand whether the highest result reflects a meaningful degree of similarity. If we do not know the threshold above which similarity is meaningful then we may not be certain if the best classification is an appropriate match. If we do know the threshold, we can choose to only consider similarities that are above this threshold. Thus, it is valuable to understand the relationship between the results of similarity measures and statistical comparisons that are used to understand the significance of results.

We provide a primary focus on *dissimilarity*, referring to *similarity* only in second instance. The reason for this is that statistical tests determine if there is a significant *difference* (rather than *similarity*) between two sets. Therefore, to provide systematic alignment between both approaches, in this paper, similarity measures are translated into measures of dissimilarity to enable a meaningful comparison with statistical tests. When dissimilarity is found to be significant, the two sets are unlikely to be identical and more likely to be dissimilar. When dissimilarity is non-significant, the difference between groups falls entirely within the acceptable margin of error of the statistical test. Such a finding indicates a higher degree of similarity than when a significant difference is found (assuming equivalent statistical power between comparisons). Further, a non-significant difference can indicate an objectively high degree of similarity given sufficient statistical power. Therefore, although the statistical significance of dissimilarity between two sets does not readily translate into a measure of significant similarity, we propose it is a useful indicator for the purposes of defining whether *similarity* is likely to be meaningful.

We use two different statistical tests to find the significance of dissimilarity. Permutation testing and the Kolmogorov-Smirnov (KS) test were chosen because they make no assumptions about the distribution of the underlying data, enabling us to explore dissimilarity between non-convex fuzzy sets. Whereas many other common statistical tests assume the data sets (or in our case, membership functions) are normally distributed. The two tests take fundamentally different approaches to compare sets, enabling us to explore how the choice of statistical test affects the resulting thresholds of significant dissimilarity.

First, Section II provides a background on fuzzy sets, dis/similarity measures and statistical tests. Next, Section III provides an overview of the synthetic data generated for the experiments and introduces two methods of using statistical tests to compare fuzzy sets. Following this, Section IV determines the significance of the results from dissimilarity measures, and Section V compares permutation testing and the KS-test on evaluating thresholds. Finally, Section VI presents conclusions.

II. Background

In this section, we present the necessary background on fuzzy sets and methods of measuring their dissimilarity, followed by an introduction of statistical tests.

A. Similarity and Dissimilarity

A discrete fuzzy set A can be expressed as a set of ordered pairs by

$$A = \{(x, \mu_A(x)) \mid x \in X\}, \tag{1}$$

where $\mu_A(x) \in [0,1]$ indicates the membership grade of the element x in the fuzzy set A, and X is a set containing all elements that may be in the fuzzy set.

Similarity and distance/dissimilarity measures have been applied to fuzzy sets in numerous applications, including pattern recognition [5], image recognition [6], fuzzy rule-base simplification [8, 9] and risk analysis [10, 11].

Let $\mathscr{P}(U)$ be the family of all crisp subsets in the universe U, and let $\mathscr{F}(U)$ be the family of all fuzzy subsets in U. A similarity measure s is a function that is most commonly given as $s : \mathscr{F}(U) \times \mathscr{F}(U) \to [0,1]$ [17–21], although some of the literature defines s as $s : \mathscr{F}(U) \times \mathscr{F}(U) \to \mathbb{R}^+$ [22, 23]. Key properties of similarity measures include [19–21, 23–29]

S1 $0 \leq s(A,B) \leq 1 \ \forall A, B \in \mathscr{F}(U)$
S2 $s(A,B) = 1 \Leftrightarrow A = B \ \forall A, B \in \mathscr{F}(U)$
S3 $s(A,B) = s(B,A) \ \forall A, B \in \mathscr{F}(U)$

S4 If $A \subseteq B \subseteq C \Rightarrow s(A,C) \leq s(A,B)$ and $s(A,C) \leq s(B,C)\ \forall A,B,C \in \mathscr{F}(U)$
S5 $s(D,D^c) = 0\ \forall D \in \mathscr{P}(U)$
S6 $s(C,C) = \max_{A,B \in \mathscr{F}(U)} s(A,B) \forall C \in \mathscr{F}(U)$
S7 $s(A,B) = 0 \Leftrightarrow A \cap B = \emptyset$

Note that the term *similarity* is loosely defined, and it is therefore not necessary (or possible) for a similarity measure to have all these properties. For example, S7 is not always included as it may be too strict [30]. Instead, S5 is often used, restricting the result of $s = 0$ to crisp sets [17, 22, 23]. However, several methods, such as Jaccard, have property S7. It has also been discussed that there are situations in which symmetry (S3) does not need to be satisfied [31]. However, a similarity measure that is not symmetrical is more generally referred to as a measure of subsethood. It also has been argued whether transitivity (S4) is necessary or even useful in some contexts [32, 33]. S6 is an alternative to S1, not limiting the similarity of identical sets to 1 [22]. A measure s that follows S6 gives a result in $\mathbb{R}^+$, whereas a measure that follows S1 gives a result in $[0,1]$.

A dissimilarity measure d is a function that is most commonly given as $d : \mathscr{F}(U) \times \mathscr{F}(U) \to [0,1]$ [21, 28, 34] , but may be defined as $d : \mathscr{F}(U) \times \mathscr{F}(U) \to \mathbb{R}^+$ [23] or $\to \mathbb{R}$ [35]. In this paper, we use dissimilarity as a measure of the difference between the fuzzy sets in membership axis, where the result is given in $[0,1]$. However, the term dissimilarity has also been used to refer to the distance in terms of the ordering on the x-axis, where the result is given in $\mathbb{R}^+$ or $\mathbb{R}$ [36].

Key properties of dissimilarity measures include [21, 23, 24, 28, 34, 35]

D1 $0 \leq d(A,B) \leq 1\ \forall A,B \in \mathscr{F}(U)$
D2 $d(A,B) = 0 \Leftrightarrow A = B\ \forall A,B \in \mathscr{F}(U)$
D3 $d(A,B) = d(B,A)\ \forall A,B \in \mathscr{F}(U)$
D4 If $A \subseteq B \subseteq C \Rightarrow d(A,C) \geq d(A,B)$ and $d(A,C) \geq d(B,C)\ \forall A,B,C \in \mathscr{F}(U)$
D5 $d(D,D^c) = \max_{A,B \in \mathscr{F}(U)} d(A,B) \forall D \in \mathscr{P}(U)$
D6 $d(A,B) = 1 \Leftrightarrow A \cap B = \emptyset\ \forall A,B \in \mathscr{F}(U)$

S1 is the counterpart to D1, S2 to D2, and so on [24] until S5 and D5. Property S6 is a counterpart to D2 if $s \to \mathbb{R}$ and $d \to \mathbb{R}$, whereas S2 in a counterpart to D2 if $s \to [0,1]$ and $d \to [0,1]$. Also note that D6 is new to this paper as a property of a dissimilarity measure that is the negation of a similarity measure with S7.

Similarity is sometimes based on the negation of difference [17, 36, 37] or dissimilarity as the negation of similarity [22, 38, 39], though other methods of transforming distance/dissimilarity into similarity and vice versa can be used [40]. In this paper, we translate a similarity measure into a dissimilarity measure using negation to provide a meaningful comparison of the measure with statistical tests. We can then translate this result into a threshold of meaningful similarity. Note that dissimilarity is not always regarded as the opposite of similarity because the features of importance may differ depending on which is being measured [31]. However, for the purposes of finding thresholds of meaningful similarity, we propose that the opposite of dissimilarity is sufficient. This is because dissimilarity is only used to determine meaningful similarity, and so the features of importance remain the same.

In this paper, we the use the Jaccard similarity measure [41] to demonstrate our method for determining thresholds of significant dissimilarity. We then compare the results of Jaccard with other measures of similarity. Jaccard's approach has properties S1, S2, S3, S4 and S7 and, for two fuzzy sets A and B on X, their similarity is

$$s_J(A,B) = \frac{\sum_{i=1}^{n} \min(\mu_A(x_i), \mu_B(x_i))}{\sum_{i=1}^{n} \max(\mu_A(x_i), \mu_B(x_i))}, \tag{2}$$

where n is the total number of discretisations within X. For dissimilarity, we use the complement of similarity, i.e.,

$$d_J = 1 - s_J \tag{3}$$

Therefore d_J as properties D1, D2, D3, D4 and D6.

B. Statistical Comparisons

We first describe the permutation test and KS-test on crisp sets, and discuss our method of extending these methods to fuzzy sets in section III.

1) Permutation Tests: We use permutation tests to determine if two data sets differ significantly. Our null hypothesis is that the variables/fuzzy-sets are the same. Consider two crisp sets A and B each with a total of n numeric values. First, we calculate the measure being tested (in our case we calculate their dissimilarity). All results on permutations of the sets are compared against this result. To create a permutation, the values within both sets are put into one pool $A \cup B$ that contains all $2n$ values. Next, a new random permutation of the two sets A' and B' is created. This is achieved by removing one value from the pool at a time (without replacement) and placing it alternatively into A' and B' until the pool is empty. We then calculate the measure we are testing (dissimilarity) on A' and B', observing if the result is higher or lower than the result on the original variables A and B.

We do this for at least 1000 random permutations and measure the percentage of times (p) the measurement on the permuted sets was higher than on the original sets. If $p > \alpha$, we conclude the difference between the variables is non-significant and accept the null hypothesis. If $p < \alpha$, we reject the null hypothesis and can state that they are found to be significantly different.

2) Kolmogorov-Smirnov Test: The KS-test is used to determine if two data sets differ significantly. It is useful for the experiments in this paper as it makes no assumptions about the distribution of the data, enabling us to evaluate skewed and multi-modal distributions. The test calculates the maximum vertical distance between the empirical distribution functions (ECDFs) of the samples. This is referred to as the D-statistic.

For example, Fig. 3 shows two sets and their ECDFS. The maximum distance D between the ECDFS is highlighted. Using D, the KS-test calculates a p-value based on the D-statistic and the distribution and sample size of the data. As with the permutation test, the null hypothesis that the samples were drawn from the same distribution is rejected if the p-value is less than the α-criterion.

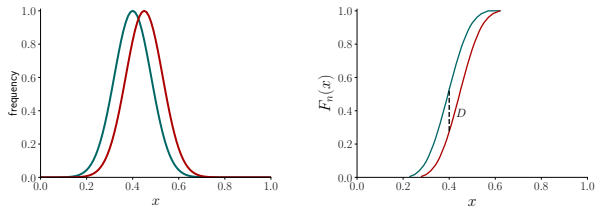


Fig. 3. Example of (a) two data sets and (b) their empirical distribution functions and their maximum vertical distance D.

III. Methods

This paper analyses the results of dissimilarity measures on 900 synthetic fuzzy sets with different types of membership functions (parametric and non-parametric). We compare results of permutation testing and the KS-test to the dissimilarity results to determine when the dissimilarity between fuzzy sets is found to be significant. In this section, we discuss how fuzzy sets were generated for the experiments, followed by a brief discussion of the chosen discretisation used in calculations. After this, we propose two methods of performing statistical tests on fuzzy sets.

A. Generating Synthetic Fuzzy Sets and Data

In this paper, we conduct experiments using three synthetic data sets containing different types of membership functions. This enables us to see if different types of membership functions affect the results of the dissimilarity measures. The generated sets are:

- L_1: Gaussian sets (e.g., Fig. 4).
- L_2: Skewed, bimodal and Gaussian sets (e.g., Fig. 5).
- L_3: Examples of fuzzy logic system outputs based on Gaussian (membership) functions (e.g., Fig. 6).

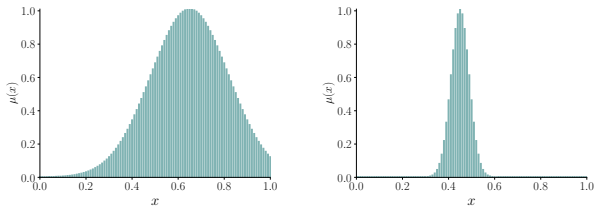
For each set, 300 synthetic fuzzy sets were created. This amount was chosen to ensure accurate conclusions by comparing a large number of fuzzy and permutation test results.

In L_1, each fuzzy set is modelled by a Gaussian membership function with a randomly chosen mean in $(0, 1)$ and a randomly chosen standard deviation in $(0.01, 0.25)$.

For each fuzzy set of L_2, two Gaussian membership functions are generated (in the same manner as for L_1) and their union is used to create sets that may be bimodal or unimodal and skewed. Note, also, that some resulting fuzzy sets are themselves Gaussian because, of the two generated functions, one is sometimes a complete subset of the other.

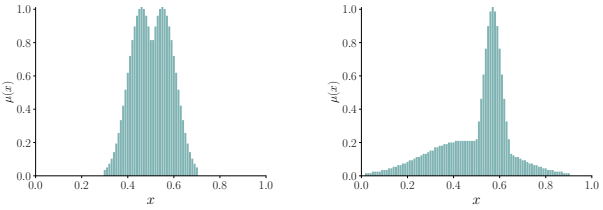
L_3 models possible outputs of a fuzzy logic system. These are based on equidistant Gaussian functions (shown in Fig. 7) representing consequents of fuzzy rules. For each fuzzy set of L_3, two or three output fuzzy sets are randomly chosen, each with a random firing level to model the result of max-min inference. Fig. 6 shows two examples of normalised fuzzy sets of L_3.

The $L1$ sets were chosen as most uses of fuzzy sets in the literature use normal, convex membership functions - the $L1$ sets reflect such cases. The $L2$ sets are designed to reflect fuzzy sets that model data that may be bi-modal or heavily skewed. The $L3$ sets are designed to aid in understanding the similarity between fuzzy logic system outputs.



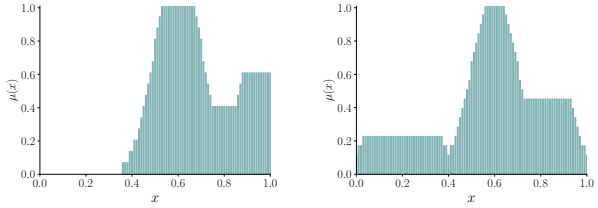
(a) (b)

Fig. 4. Two normally distributed, synthetic data sets.



(a) (b)

Fig. 5. Two non-normally distributed, synthetic data sets.



(a) (b)

Fig. 6. Two synthetic data sets modelling a fuzzy logic system output.

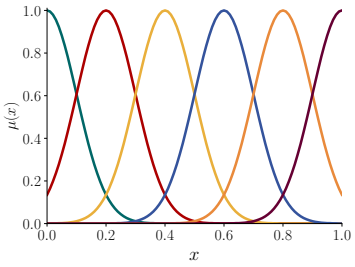
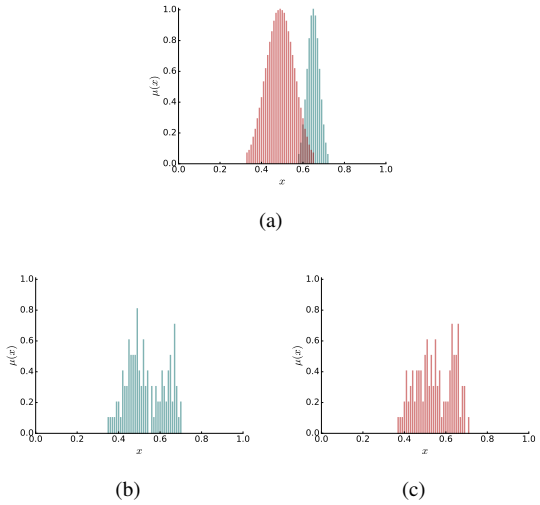


Fig. 7. Output fuzzy sets used to generate fuzzy sets of L_3.

B. Choosing Appropriate Discretisation

Fuzzy sets, such as those from the output of a fuzzy logic system, are commonly not modelled as continuous functions. For example, in order to assess their dissimilarity, the universe of discourse is discretised. It is therefore necessary to establish an appropriate level of discretisation to ensure accurate results. We first conduct experiments to observe the effects of using different levels of discretisation. The chosen ideal level of



(a) (b) (c)

Fig. 8. Example of two fuzzy sets and a random permutation.

discretisation is then used to determine the threshold that indicates significant results in dissimilarity measures.

In the Appendix, we present experiments on the results of dissimilarity at different levels of discretisation to select an appropriate level. Based on the results in Table IV and Fig. 19, we suggest using precision at every 1% along X is sufficient for most applications. In our case, $X = \{0, 0.01, 0.02, ...0.99, 1.0\}$ Note that although L_3 shows a greater change in accuracy at higher precision than L_1 or L_2, we still set precision at 1% as sufficient.

C. Permutation Testing on Fuzzy Sets

To permute fuzzy sets, we deconstruct them into crisp sets, permute the crisp sets and then reconstruct the permutation into fuzzy sets. We achieve this by discretising the fuzzy sets along the x and y axes into n and m discrete points, respectively. The total occurrences of a value x in the crisp set is $m \cdot \mu(x)$; we use $m = 10$ to ensure an integer result. For example, if $A = \{(x, \mu(x)) \mid \forall x \in X\} = \{(0.3, 0.5), (0.4, 1)\}$, then the crisp set of A (denoted A^c) will have 5 occurrences of the value 0.3 and 10 occurrences of the value 0.4. After converting two fuzzy sets (A and B) into two crisp sets (A^c and B^c), we can then pool the crisp sets and create a random permutation (A'^c and B'^c) as normal. The results of the permutation are then constructed as fuzzy sets. The membership of x in the new fuzzy set A' is the total occurrences of x in the crisp set A'^c divided by m. When creating a permutation, no value is allocated to a set more than m times to ensure a membership $\mu(x) > 1$ does not occur. Fig. 8 shows an example of two fuzzy sets and a random permutation.

From here, the method of permutation testing is the same as for crisp sets. We measure the dissimilarity of the permuted sets and compare if this is higher than the dissimilarity for the original sets. This is done for 1000 permutations and the percentage of times the dissimilarity was higher is our p-value.

The α-criterion 0.01 is used, with the null hypothesis that two sets are drawn from the same distribution. Therefore, if a pair of fuzzy sets have a p-value below 0.01, the corresponding dissimilarity result calculated with the measure is also considered significant. Note that we discuss using different α-criteria in section IV-B.

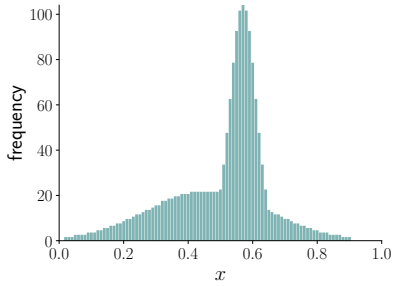


Fig. 9. The crisp set derived from the fuzzy set in Fig. 5(b).

As an example, consider two fuzzy sets that are identical - their dissimilarity is 0. All other permutations of these two fuzzy sets will result in non-identical sets with a dissimilarity greater than 0. Therefore, 100% of the permuted fuzzy sets will have a greater dissimilarity than the original, resulting in a p-value of 1, meaning we accept the null hypothesis that the sets are the same. As another example, consider two fuzzy sets that are disjoint - their dissimilarity is 1. All other permutations of these two fuzzy sets will not be disjoint and have a dissimilarity less than 1. Therefore, 0% of the permuted fuzzy sets will have a greater dissimilarity than the original, resulting in a p-value of 0, meaning we reject the null hypothesis and conclude the sets are significantly different.

D. Kolmogorov-Smirnov Testing on Fuzzy Sets

The KS-test is designed to compare crisp sets rather than fuzzy sets. Therefore, we must convert the fuzzy sets into appropriate crisp sets. This is achieved using the same method as described for permutation testing. That is, we divide the x and y axes into n and m discrete points and create a crisp set where the number of times x occurs is equal to $m \cdot \mu(x)$. We use this process to convert each fuzzy set into a crisp set. As an example, Fig. 9 shows the crisp set derived from the fuzzy set in Fig. 5(b).

We use the KS-test to test the null hypothesis that the crisp sets were drawn from the same distribution. We then compare the p-value from the KS-test (performed on the crisp sets) with the dissimilarity result (performed on the fuzzy sets). We then determine a threshold for the dissimilarity measure using the α-criterion 0.01. If a pair of fuzzy sets have a p-value below 0.01, their dissimilarity result according to the fuzzy dissimilarity measure is considered significant.

IV. The Significance of Dissimilarity

A. Finding Thresholds of Significant Dissimilarity

We wish to determine thresholds that indicate when the similarity between fuzzy sets is noteworthy or otherwise. To achieve this, we first find thresholds for the dissimilarity measure, where values above this threshold indicate a statistically

significant degree of dissimilarity for a given α-criterion, and values below are non-significant. We first demonstrate this method (which can be applied to any (dis)similarity measure) with Jaccard (3), providing results for additional measures in Section IV-D.

Within each set (L_1, L_2 and L_3), all 300 fuzzy sets are compared with themselves and with each other fuzzy set. This makes for a total of 45150 comparisons with each method in each set. Note, as every fuzzy set is compared with itself, there are at least 300 identical pairs compared each in L_1, L_2 and L_3. We then compare the Jaccard results against the statistical test results.

Our null hypothesis is that both samples are drawn from the same distribution, i.e. that the fuzzy sets are the same. If the p-value from the statistical test is less than our α-criterion of 0.01 then we reject this hypothesis. Thus, we propose a p-value below 0.01 for a pair of fuzzy sets provides evidence that their dissimilarity is significant. It follows that for pairs with a p-value higher than 0.01, their fuzzy set dissimilarity is non-significant.

Intuitively, from the point of view of the dissimilarity results, we expect the existence of a switch-point or switch-range σ where the results cross the statistical significance/non-significance boundary. The values returned by a dissimilarity measure d can be categorised as follows: $d(A, B) \in [[0, \sigma), [\sigma, 1]]$, where σ is either a real-valued number or a real-valued interval in $[0, 1]$ determined by the α-criterion 0.01.

Fig. 10 show scatter plots of the Jaccard results for all paired fuzzy set combinations in all sets (L_1, L_2 and L_3), compared against the permutation test results. Fig. 11 show the same for the KS-test. Each point in the plots captures the statistical test p-value (x-axis) against the dissimilarity measure result (y-axis) for a pair of fuzzy sets.

To help focus on the pairs that are below and above the threshold of statistical significance, Fig. 12 shows a subset of the results for d_J on L_2 (from Fig. 10(b)) with the range of p-values in $[0, 0.1]$. We have drawn a red vertical line at the α-criterion 0.01 and orange dashed lines to highlight thresholds of significance, which we derive next.

In Fig. 12, where $p < 0.01$ the range of values from d_J (3) is wide. Specifically

$$d_J(L_{2_i}, L_{2_j}) \in [0.2008, 1.0] \mid p(L_{2_i}, L_{2_j}) < 0.01, \\ i, j \in \{0, 1, ..., 299\},$$

where $p(L_{2_i}, L_{2_j})$ refers to the resulting p-value from the permutation test when comparing fuzzy sets L_{2_i} and L_{2_j}.

Where $p \geq 0.01$, the range of (3) is

$$d_J(L_{2_i}, L_{2_j}) \in [0.0, 0.3263] \mid p(L_{2_i}, L_{2_j}) \geq 0.01, \\ i, j \in \{0, 1, ..., 299\}.$$

Note that this leaves an interval of overlap at $d_J(L_{2_i}, L_{2_j}) \in [0.2008, 0.3263)$ where $p(L_{2_i}, L_{2_j})$ may be either less than or greater than the α-criterion 0.01. Table I shows these intervals of overlap for (3) on all sets (L_1, L_2, and L_3) for permutation testing and the KS test.

We set strong and weak thresholds that determine the worth of the dissimilarity results. Continuing the same example (observing the results of (3) in L_2 for the permutation test), we approximate the interval of results that may be significant or non-significant to $[0.2, 0.35]$. We set a strong threshold at 0.35 as all pairs where $d_J \geq 0.35$ were found to be significantly dissimilar. We set a weak threshold at 0.2 as most pairs where $d_J \geq 0.2$ were found to be significantly dissimilar, but some were non-significant. Where $d_J < 0.2$, no pairs were found to be significantly dissimilar.

TABLE I
VALUES OF DISSIMILARITY ABOVE WHICH THE RESULT OF d_J (3) WAS ALWAYS FOUND TO BE SIGNIFICANT (t_s) AND VALUES BELOW WHICH DISSIMILARITY WAS ALWAYS FOUND TO BE NON-SIGNIFICANT (t_w) USING PERMUTATION TESTING AND THE KS-TEST. BETWEEN t_s AND t_w SIGNIFICANCE VARIED.

perm	L_1	L_2	L_3
t_s	0.3501	0.3263	0.7939
t_w	0.2011	0.2008	0.3338
KS	L_1	L_2	L_3
t_s	0.2725	0.3922	0.4862
t_w	0.1242	0.0911	0.106

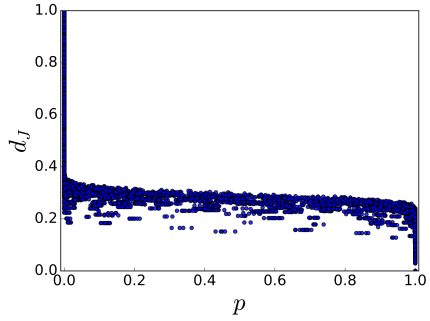
According to the permutation test, the strong and weak thresholds are approximately the same for groups L_1 and L_2, but both thresholds increase for group L_3. Using the the KS-test, the strong threshold increases as the fuzzy sets increase in complexity (e.g., in this case, are formed by the union of multiple functions instead of a single function), but the weak thresholds remain approximately the same.

To determine thresholds for similarity measures we use the complement of the dissimilarity thresholds and swap the threshold labels. For the permutation results on L_2, the strong threshold of S_J is 0.8. Above 0.8, similarity is likely to be meaningful (dissimilarity was always found to be non-significant). The weak threshold is 0.65 as pairs above this may or may not have meaningful similarity, with a high number of both significant and non-significant dissimilarity. Below the weak threshold, similarity is unlikely to be meaningful as dissimilarity was always found to be significant. Fig. 13 demonstrates how the thresholds are flipped from dissimilarity (Fig. 12) to similarity.

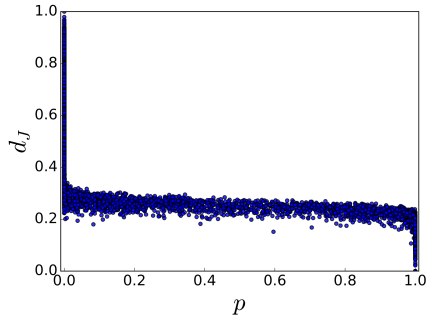
Table II summarises the weak and strong thresholds of dissimilarity (3) and similarity (2) on each set (L_1, L_2, and L_3) using permutation testing and the KS-test. A value of similarity or dissimilarity above the strong threshold is likely to be meaningful and below the weak threshold is unlikely to be meaningful. Values between the weak and strong thresholds have uncertain worth.

B. Choosing a Different α-criterion

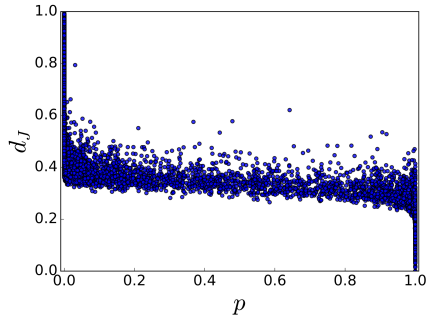
The most appropriate α-criterion can vary between tests and situations. In different cases it may be more important to minimise Type I or Type II errors respectively. Moreover, if multiple comparisons are conducted then it will be necessary to adjust for the resulting α-inflation, in order to maintain a suitable family-wise error rate. Fig. 14 shows thresholds corresponding to a range of different α-criteria, according to both statistical tests on (3) for each fuzzy set group. This



(a) L_1

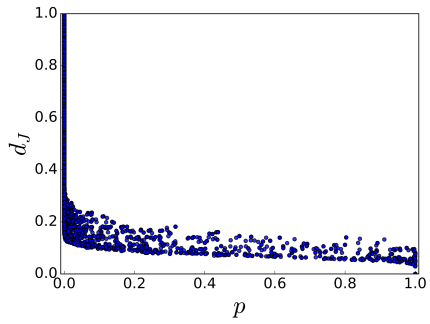


(b) L_2

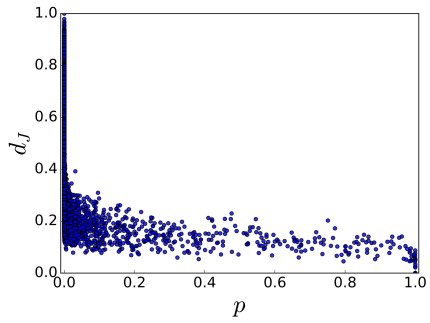


(c) L_3

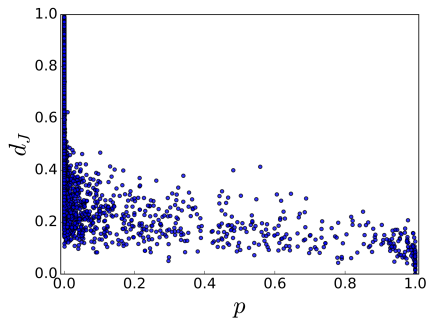
Fig. 10. Results of the Jaccard dissimilarity measure against the p-values of the permutation test on the three different sets of synthetic data.



(a) L_1



(b) L_2



(c) L_3

Fig. 11. Results of the Jaccard dissimilarity measure against the p-values of the KS-test on the three different sets of synthetic data.

illustration of the increase in similarity thresholds, according to the decrease in α-criterion, also indicates the extent to which the thresholds require modification in cases of multiple comparisons.

For group $L3$, there are several pairs from the permutation test that may be considered outliers and so the strong threshold may be higher than necessary. The strong threshold also quickly drops for a higher α-criterion. The thresholds of group L_3 are, however, consistently higher than those for L_1 and L_2. The thresholds for groups L_1 and L_2 for both statistical tests remain stable across different α-criteria. That is, the choice of α-criterion for L_1 and L_2 sets has no strong effect on the resulting thresholds, whereas for L_3 sets, a strong effect was found. In addition, according to the KS-test the threshold drops quicker as the α-criterion is increased than according to the permutation test. This may be a result of the wider area

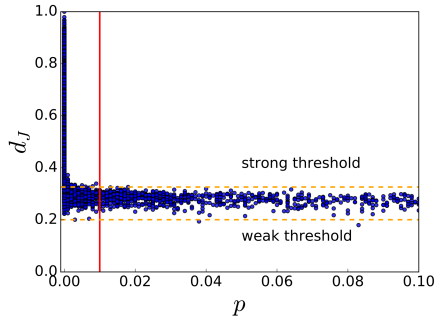


Fig. 12. A subset of the results (where $p < 0.1$) of the Jaccard dissimilarity measure against the p-values of the permutation test on L_2. The red line highlights the α-criterion of 0.01. The orange dashed lines indicate the strong threshold (largest value of (3) where $p > 0.1$) and the weak threshold (smallest value of (3) where $p < 0.1$).

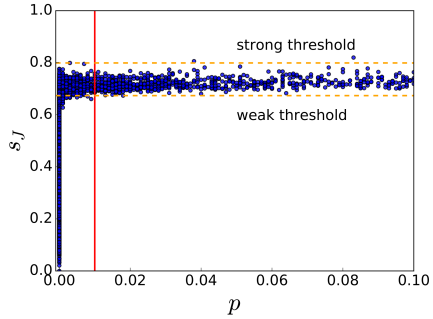


Fig. 13. A subset of the results (where $p < 0.1$) of the Jaccard similarity measure against the p-values of the permutation test on L_2. The red line highlights the α-criterion of 0.01. The orange dashed lines indicate the strong threshold (largest value of (3) where $p < 0.1$ and the weak threshold (smallest value of (3) where $p > 0.1$).

TABLE II
APPROXIMATED WEAK AND STRONG THRESHOLDS OF DISSIMILARITY (3) AND SIMILARITY (2) ON EACH SET USING PERMUTATION TESTING AND THE KS-TEST.

perm.	Strength	L_1	L_2	L_3
d_J (3)	strong	0.35	0.35	0.8
	weak	0.2	0.2	0.35
s_J (2)	strong	0.8	0.8	0.65
	weak	0.65	0.65	0.2

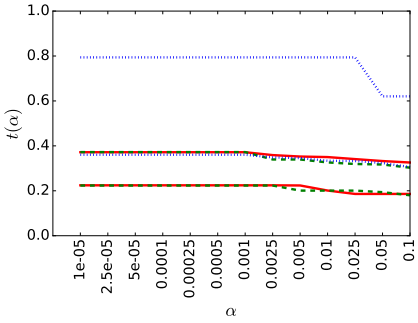
KS.	Strength	L_1	L_2	L_3
d_J (3)	weak	0.1	0.1	0.1
	strong	0.3	0.4	0.5
s_J (2)	weak	0.7	0.6	0.5
	strong	0.9	0.9	0.9

of uncertainty (the difference between the strong and weak thresholds) provided by the KS-test.

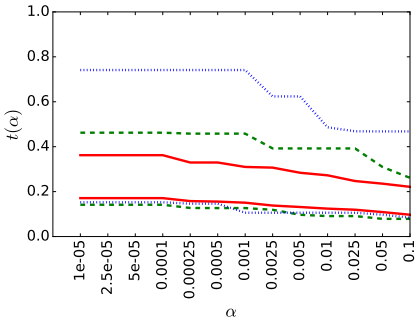
C. The Effect of Sample Size

Consider a statistical test calculated for two data sets to test the null hypothesis that they are drawn from the same distribution. The distance between the sets according to the test is x and this is assigned a p-value less than our alpha-criterion; that is, the sets are found to be significantly different. Next, consider we use the same test on a subset of the data. The test finds the same difference x but assigns a p-value greater than our alpha-criterion; that is, the sets are not significantly different. In these two examples, different p-values were assigned due to the different sample sizes. In the second case, the difference between the subsets is more likely to be due to the small sample size chosen than because the sets are from different distributions. In this case, a larger difference between the sets is needed to confidently state they are drawn from different distributions.

Likewise, if we use a subset of the fuzzy sets and measure their dissimilarity, we expect to need a larger degree of dissimilarity to conclude the sets are significantly different. That is, the fewer discretisations used to compare the fuzzy sets, the greater the dissimilarity must be for the evidence to be strong enough to conclude they are significantly different.



(a)



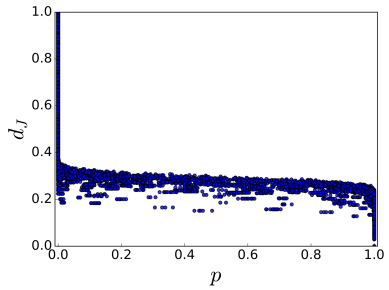
(b)

Fig. 14. Strong (high) and weak (low) thresholds found for d_J (3) by (a) permutation testing and (b) KS test for groups L_1 (red solid), L_2 (green dashed) and L_3 (blue dotted).

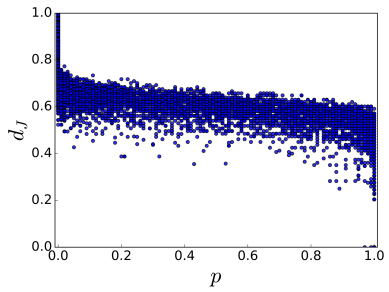
To test this hypothesis, we run the permutation test on subsets of the fuzzy sets. In the previous tests, we used approximately 200 data points to define the fuzzy sets. We now reduce these to random samples of 100 data points and 50 data points, from which we construct fuzzy sets. We expect the threshold of significant dissimilarity to increase as the number of data points used to construct the sets decreases.

Fig. 15 shows the results for d_J (3) with the permutation test using the full data set and using subsets. The figure shows a clear effect of threshold on sample size. As the sample size decreases, the threshold increases. In addition, the area of uncertainty (that is, the gap between the weak and strong thresholds) also increases. This empirically confirms our hypothesis that when less information is used to construct fuzzy sets, a greater degree of dissimilarity between them is needed to confidently conclude they are from different distributions.

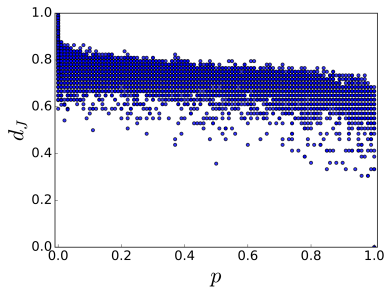
Fig. 16 show results on different sample sizes when using the KS-test. As with permutation testing, a smaller sample size increases the threshold for significant dissimilarity. However, these results also show that the KS-test is sensitive to sample size, providing a more discrete range of p-values. This suggests that permutation testing may be more suitable to finding thresholds of significant dissimilarity, particularly when a small sample size (i.e. small range of discretisations) is used. The same trend of results is found with other dissimilarity measures.



(a) n =all (≈ 200)

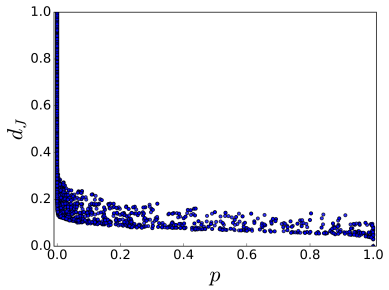


(b) $n = 100$

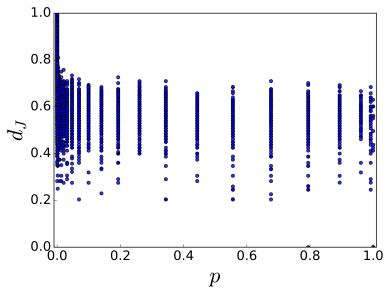


(c) $n = 50$

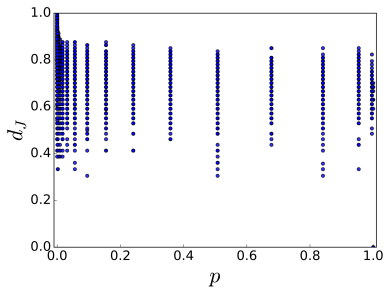
Fig. 15. Results of the permutation test and the Jaccard dissimilarity on group 1 (Gaussian sets) with different sample sizes (n).



(a) n =all (≈ 200)



(b) $n = 100$



(c) $n = 50$

Fig. 16. Results of the KS-test and the Jaccard dissimilarity on group 1 (Gaussian sets) with different sample sizes (n).

D. Comparisons of Different Similarity Measures

Next, we provide the threshold results for other dissimilarity measures for comparison. In future, the process of deriving the thresholds put forward in this paper can also be used to generate further results for additional measures. The measures for which we include results are the Sørensen-Dice coefficient [42]:

$$d_D(A,B) = 1 - \frac{2\sum_{i=1}^{n} \min(\mu_A(x_i), \mu_B(x_i))}{\sum_{i=1}^{n} \mu_A(x_i) + \sum_{i=1}^{n} \mu_B(x_i)}. \quad (4)$$

two measures by Pappis and Karacapilidis [17]:

$$d_{P1}(A,B) = \max_i |\mu_A(x_i) - \mu_B(x_i)| \quad (5)$$

$$d_{P2}(A,B) = \frac{\sum_{i=1}^{n}(|\mu_A(x_i) - \mu_B(x_i)|)}{\sum_{i=1}^{n}(\mu_A(x_i) + \mu_B(x_i))} \quad (6)$$

by Chen [43]:

$$d_C(A,B) = 1 - \frac{\sum_{i=1}^{n} \mu_A(x_i) \cdot \mu_A(x_i)}{\max\{\sum_{i=1}^{n} \mu_A(x_i)^2, \sum_{i=1}^{n} \mu_A(x_i)^2\}} \quad (7)$$

and by Zwick et al. [36]:

$$d_Z(A,B) = 1 - \max_i \mu_{A\cap B}(x_i) \quad (8)$$

Fig. 17 shows the results of d_J (3) and measures (4) - (8) using permutation testing applied to the $L1$ sets; results for $L2$ and $L3$ are shown in the Appendix. Table III shows the numerical thresholds of the dissimilarity measures for all three sets ($L1$, $L2$, $L3$) and the translated thresholds (the complement of the dissimilarity thresholds) for the related similarity measures. Figs. 17(a), (b), (d) and (e) show similar results as the methods have the same properties (D1, D2, D3, D4 and D6). Fig. 17(c) (measure (5)) and Fig. 17(f) (measure (8)) have noticeably different results as their properties differ.

The measure (5) (Fig. 17(c)) has the properties D1, D2, D3 and D6. Unlike the other measures, it does not have D4 and, instead, $d_{P1}(A,B) = 1 \quad \exists x\ (\mu_A(x) = 1 \wedge \mu_B(x) = 0) \vee (\mu_A(x) = 0 \wedge \mu_B(x) = 1)$. Therefore, unlike the other measures, a dissimilarity of 1 may be given even if the fuzzy sets overlap. Likewise, the dissimilarity is large even if the

TABLE III
APPROXIMATE WEAK AND STRONG THRESHOLDS OF DIFFERENT DISSIMILARITY AND SIMILARITY MEASURES ON EACH SET USING THE PERMUTATION TEST. AN * INDICATES $s = 1 - d$, WHERE d IS THE GIVEN DISSIMILARITY EQUATION.

	Strength	L_1	L_2	L_3
d_J (3)	strong	0.35	0.35	0.8
	weak	0.2	0.2	0.35
d_D (4)	strong	0.2	0.2	0.65
	weak	0.1	0.1	0.2
d_{P1} (5)	weak	0.8	0.75	0.78
	strong	0.98	0.99	0.95
d_{P2} (6)	weak	0.2	0.2	0.15
	strong	0.4	0.6	0.3
d_C (7)	weak	0.1	0.1	0.1
	strong	0.3	0.4	0.25
d_Z (8)	weak	0.0	0.0	0.0
	strong	0.6	0.68	0.57
s_J (3)*	strong	0.8	0.8	0.65
	weak	0.65	0.65	0.2
s_D (4)*	strong	0.9	0.9	0.8
	weak	0.8	0.8	0.35
s_{P1} (5)*	weak	0.02	0.01	0.05
	strong	0.2	0.25	0.22
s_{P2} (6)*	weak	0.6	0.4	0.7
	strong	0.8	0.8	0.85
s_C (7)*	weak	0.7	0.6	0.75
	strong	0.9	0.9	0.9
s_Z (8)*	weak	0.4	0.32	0.43
	strong	1	1	1

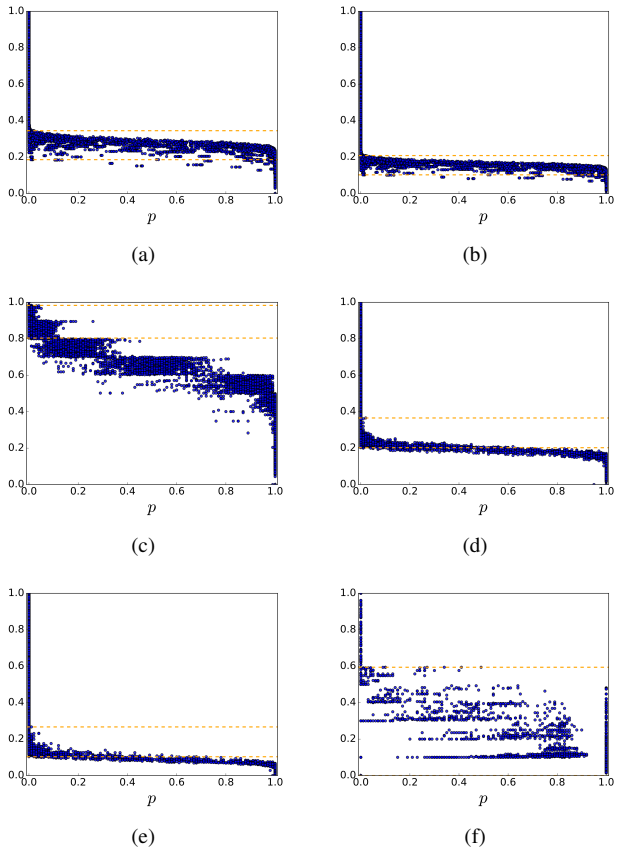


Fig. 17. Results of the permutation test on $L1$ sets for a) d_J (3), b) d_D (4), c) d_{P1} (5), d) d_{P2} (6), e) d_C (7) and f) d_Z (8). Orange dashed lines show the strong (upper) and weak (lower) thresholds of the measure. Numerical values of the thresholds are in Table III.

fuzzy sets overlap so long as there is a large difference in membership at some value of x. It is for this reason that the values of dissimilarity from (5) are much larger than with the other measures and that, subsequently, the measure has a much higher threshold.

Measure (8) (Fig. 17(f)) has the properties D1, D2, D3 and D6. The measure considers only the maximum membership of the intersection of the sets, rather than looking at the fuzzy sets as a whole. It has the property $d_Z(A, B) = 0 \quad \exists x \ \mu_A(x) = 1 \wedge \mu_B(x) = 1$ even if the sets are not identical. If then follows that d_Z gives a *low* result if the intersection of the fuzzy sets has *high* membership at any point, even if they have little overlap. This is why the threshold of (8) is larger than other methods except (5) in Fig. 17(c).

V. PERMUTATION TESTING VS KS-TEST

To perform permutation testing or the KS-test, the fuzzy sets are converted into crisp sets. Next, in permutation testing, the crisp sets are converted into fuzzy sets to calculate their dissimilarity. However, for the KS-test, the sets remain crisp in order to compare them. We found that the thresholds of significant dissimilarity differ depending on the statistical test used due to the different methods taken.

The benefit of permutation testing and the KS test, and the reason for their choice, is that they make no underlying assumptions about the distribution of the data/fuzzy sets. However, we propose that permutation testing is a more appropriate choice to find significant dissimilarity. This is because the KS-test requires a comparison of crisp sets instead of fuzzy sets, and it has been demonstrated that its results are highly affected by discretisation (sample size). The permutation test, however, does not rely on a lower level of abstraction (that is, we can run the test on the fuzzy sets themselves) and its range of p-values is not affected by sample size.

In addition, permutation testing is beneficial because it can be naturally extended to type-2 fuzzy sets. To extend the work to type-2 fuzzy sets using permutation testing, the fuzzy sets can be discretised along the x, y and z (or x, u and μ) axes. In the type-1 case, the union of the primary membership values of the fuzzy sets is redistributed into two new sets. For type-2 fuzzy sets, the union of both the primary and secondary membership values can be discretised and redistributed. Conversion of type-2 fuzzy sets for use with the KS-test would require a lower level of abstraction and potentially loss of detail about the sets.

VI. CONCLUSIONS

We propose a general method of finding thresholds of significant dissimilarity (and meaningful similarity) on fuzzy sets using statistical tests. We demonstrate this method for the Jaccard similarity measure using permutation testing and the KS-test, and compare results across several different similarity measures. As statistical comparisons of data focus on finding if two sets are significantly different, we first measure the

significance of dissimilarity measures and translate the results to understand similarity measures. Although the statistical significance of dissimilarity between two sets does not directly convert into a measure of significant similarity, we propose it as a reasonable candidate criterion for indicating whether similarity is likely to be meaningful.

We discover strong and weak thresholds that indicate the meaningfulness of dissimilarity/similarity results. When measuring dissimilarity, results above the strong threshold are found to be significant and results below the weak threshold are non-significant, falling within the margin of error of the statistical test. Results between these thresholds have uncertain significance. Given sufficient statistical power, we translate these thresholds of dissimilarity to strong and weak thresholds of similarity. Above the strong threshold we propose similarity is likely to be meaningful and above the weak threshold similarity may be meaningful. Below the weak threshold, results were consistently non-significant and therefore the similarity is unlikely to be meaningful.

We conduct experiments on different types of membership functions. $L1$) normal, convex sets $L2$) bi-modal, skewed sets, and $L3$) multi-modal sets (emulating fuzzy logic system outputs). We show that thresholds differ between simple and complex membership functions, and we show that thresholds are subject to the sample size of the data used to construct the fuzzy sets - the smaller the sample size, the lower the threshold of similarity. When measuring the similarity between fuzzy sets that are normal and convex, we suggest using the thresholds derived using $L1$ sets. If using fuzzy sets that model data that may be bi-modal or heavily skewed, we recommend using the thresholds derived using $L2$ sets. If measuring the similarity between outputs of fuzzy logic systems, we recommend using the thresholds from $L3$ sets.

The results of each method can be used in real world applications, beyond the synthetic examples in this paper. Where fuzzy sets are known to be normally distributed, the thresholds from the $L1$ sets can be used to determine an important level of similarity. If fuzzy sets are non-convex, it is best to use the thresholds from groups $L2$ or $L3$. The results from one measure should not be used to determine noteworthy significance of a different measure. This is because, as shown empirically, different measures have different thresholds. To determine thresholds of similarity measures, a general mathematical solution cannot be made. Permutation testing on a large collection of fuzzy set pairs is the most reliable method for finding thresholds of significance.

Our method of finding thresholds of importance in similarity measures is a general one and can be applied to other metric (dis)similarity measures. The code used to find thresholds (along with the numerical results) has been made available online at https://bitbucket.org/JosieMcCulloch/dissig/ and at https://lucidresearch.org/software to enable tests on other measures in the future.

APPENDIX A
CHOOSING APPROPRIATE DISCRETISATION

In this section, we conduct experiments to determine the appropriate level of discretisation in X required for accurate dissimilarity measure results. (Note that discretisation steps are always equidistant.) This will be used to ensure that a suitable level of accuracy is used when determining significant dissimilarity. Note that this section demonstrates the process using the Jaccard dissimilarity measure (3).

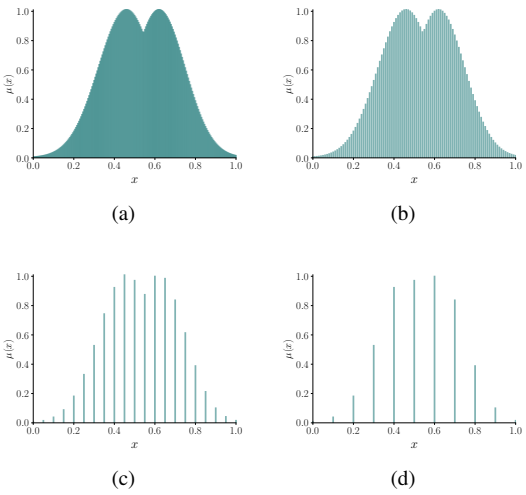


Fig. 18. An example of a non-normally distributed fuzzy set at five different levels of precision.; (a) $L^{0.5}$, (b) L^{1}, (c) L^{5}, (d) L^{10}.

TABLE IV
AVERAGE DIFFERENCE IN RESULTS OF S_J AT DIFFERENT LEVELS OF DISCRETISATION.

	L_1	L_2	L_3
$d_b^{0.1-0.5}$	0.00011	0.00018	0.00171
$d_b^{0.5-1}$	0.00024	0.00029	0.00166
d_b^{1-5}	0.00456	0.00422	0.00976
d_b^{5-10}	0.02091	0.01489	0.01493
Total compared	44726	44850	44409

The following is carried out for each set (L_1, L_2 and L_3) separately. Each fuzzy set is modelled at five different levels of precision in X. First, we model the fuzzy sets discretely with increments at every 0.1% along X. We call the resulting

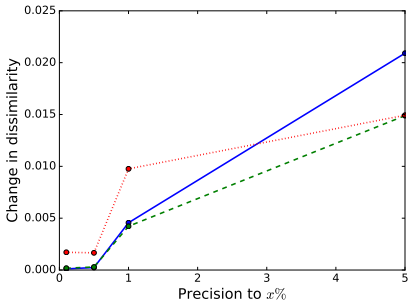


Fig. 19. Graphical representation of the difference in results of S_J at different levels of discretisation for L_1 (blue, solid), L_2 (green, dashed) and L_3 (red, dotted). As the frequency of measurements increases (at lower percentages of precision) the change in dissimilarity decreases.

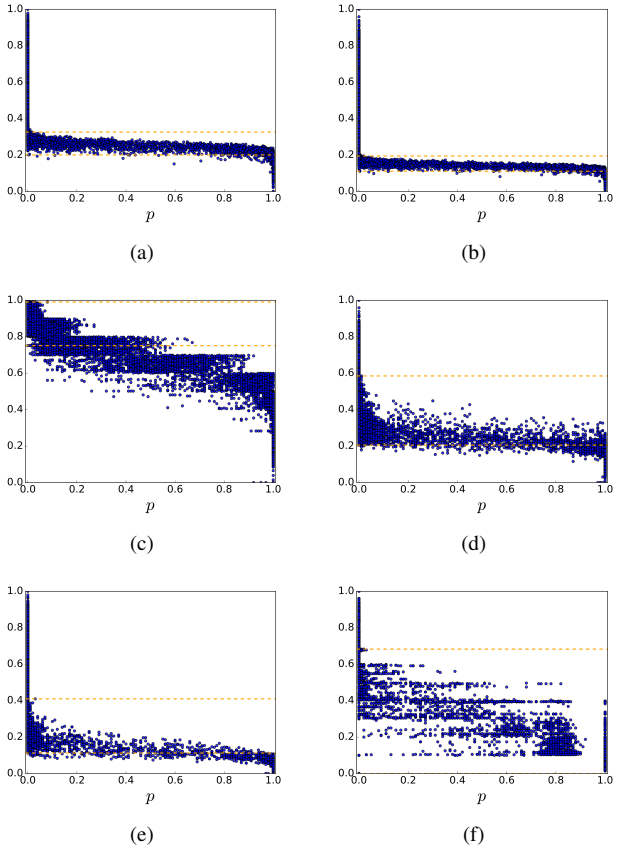


Fig. 20. Results of a) d_J, b) d_D, c) d_{P1}, d) d_{P2}, e) d_C and d_Z p-values of the permutation test on the $L2$ sets. Orange dashed lines show the strong (upper) and weak (lower) thresholds of the measure. Numerical values of the thresholds are in Table III.

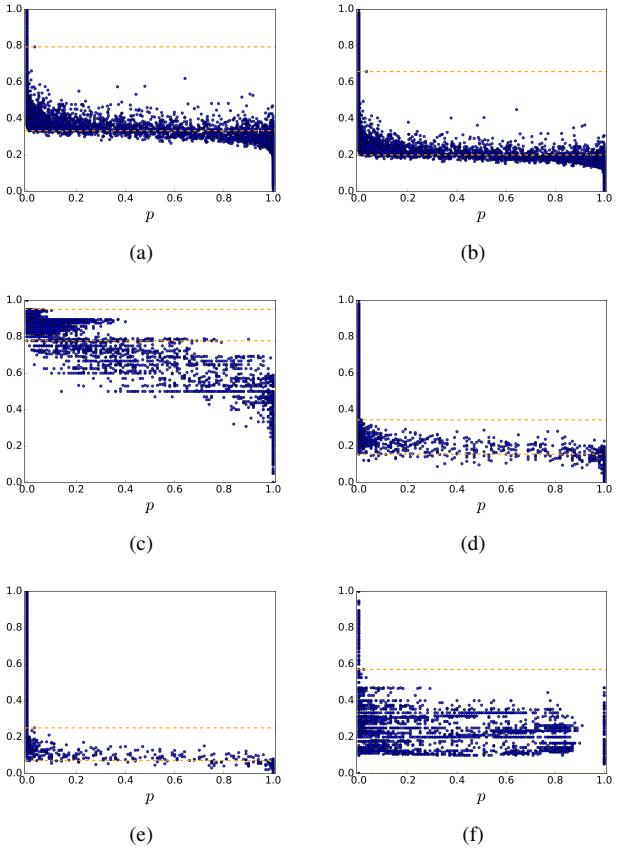


Fig. 21. Results of a) d_J, b) d_D, c) d_{P1}, d) d_{P2}, e) d_C and d_Z p-values of the permutation test on the $L3$ sets. Orange dashed lines show the strong (upper) and weak (lower) thresholds of the measure. Numerical values of the thresholds are in Table III.

set $S_b^{0.1}$, defined as follows

$$S_b^{0.1} = \{L_{b_i}^{0.1} \mid i \in \{0, 1, ..., 299\}, \\ X = [0, 0.001, 0.002, ..., 0.999, 1.0]\}$$

where $b \in \{1, 2, 3\}$ refers to the set type (i.e., L_1, L_2, or L_3), and L_{b_i} refers to the i^{th} set in L_b.

We also model the same fuzzy sets at lower degrees of precision, with X incremented every 0.5% (i.e. $X = [0, 0.005, ...0.995, 1.0]$) (denoted $L^{0.5}$), incremented every 1% (L^1), every 5% (L^5) and every 10% (L^{10}). Thus, the i^{th} fuzzy set within $L_b^{0.1}$, $L_b^{0.5}$, L_b^1, etc. is the same distribution of data modelled discretely at different degrees of precision. Fig. 18 shows an example fuzzy set at four different levels of precision. The highest level of precision ($L^{0.1}$) has been omitted from the figure as it is not possible to discern it from $L^{0.5}$ at the given resolution of the figures.

We want to determine how much the measured dissimilarity differs between the same pair of fuzzy sets at different degrees of precision. Identical and disjoint fuzzy sets will always result in a fixed value of dissimilarity (0 and 1, respectively) regardless of the precision, so we discard these from the comparison. This leaves us with a set of (i, j) pairs as follows:

$$P_b = \{(i, j) \mid 0 < d_J(L_{b_i}^{0.1}, L_{b_j}^{0.1}) < 1, \\ i, j \in \{0, 1, ..., 299\}\}$$

where b refers to the set type (L_1, L_2, or L_3). Note that the highest level of precision is used to find and exclude identical and disjoint pairs. Table IV shows how many remaining pairs are compared after identical and disjoint pairs have been removed.

We then calculate the difference in dissimilarity between each step of precision. We take the absolute difference as we are only interested in how much dissimilarity changes between levels of discretisation. It is not important if this change increases or decreases the result. The absolute differences between the dissimilarities of sets $L^{0.1}$ and of $L^{0.5}$ are

$$D_b^{0.1-0.5} = \{|d_J(L_{b_i}^{0.1}, L_{b_j}^{0.1}) - d_J(L_{b_i}^{0.5}, L_{b_j}^{0.5})| \\ \mid (i, j) \in P_b\} \quad (9)$$

The average difference is then calculated as

$$d_b^{0.1-0.5} = \frac{\sum_{d \in D_b^{0.1-0.5}} d}{n(P_b)} \quad (10)$$

where $n(P_b)$ denotes the cardinality of P_b. In the same manner as (9), we also calculate $D_b^{0.5-1}$, D_b^{1-5} and D_b^{5-10}, and their average results with (10). Table IV shows these results numerically and Fig. 19 shows them graphically.

As expected, the average difference in dissimilarity reduces as the level of precision increases. It appears that accuracy

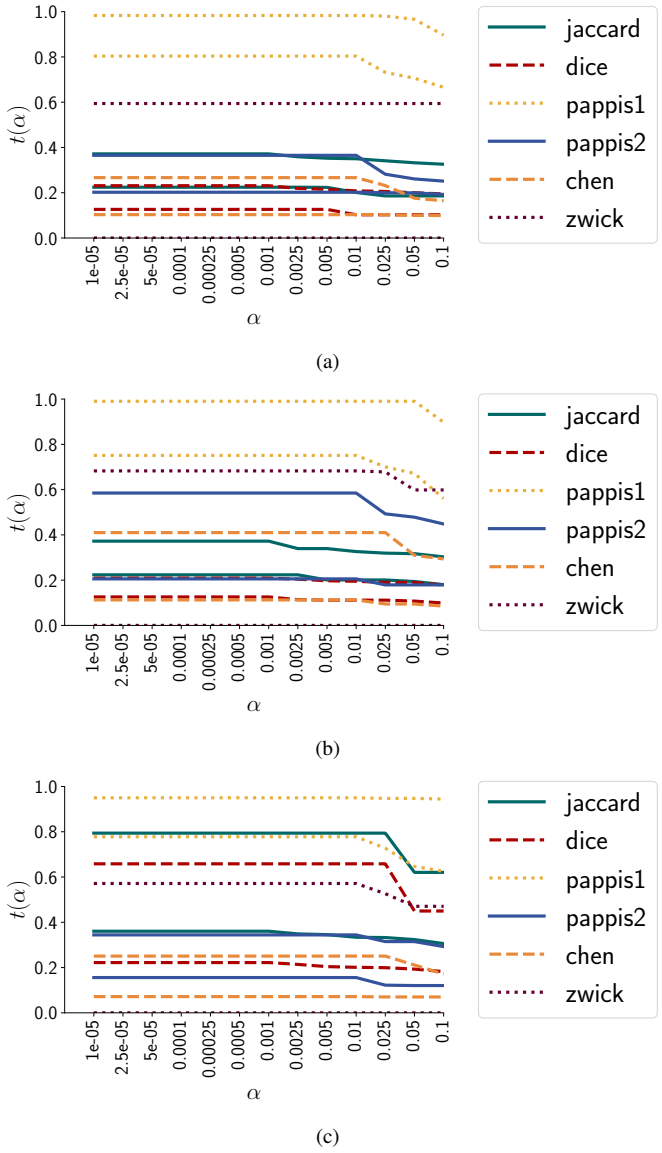


Fig. 22. Strong (high) and weak (low) thresholds found for each method by permutation testing for the dissimilarity measures for a) $L1$, b) $L2$ and c) $L3$.

in measured dissimilarity begins to plateau after precision to 1% ($S^{0.1}$). Between precision to 1% (S^{1}) and 0.5% ($S^{0.5}$), the average difference in results is only 0.00024, 0.00029 and 0.00166 for L_1, L_2 and L_3, respectively.

Based on the results in Table IV and Fig. 19, we suggest using precision at every 1% along X is sufficient for most applications. This is, of course, only a guideline. For example, in applications where fuzzy sets are considerably narrow or where X is much larger, higher precision may be required.

Appendix B
Visualisations of Dissimilarity Measure Results

Figures 20 and 21 show the results of d_J (3) and measures (4) - (8) applied to the $L2$ and $L3$ sets, respectively. Fig. 17 shows the same for $L1$ sets, and Table III shows the numerical thresholds of the dissimilarity measures for all three sets and the translated thresholds for the related similarity measures. In addition, Fig. 22 shows how the thresholds of each method compare and shows the results are consistent across different alpha-criteria.

References

[1] D. Wu and J. M. Mendel, "A vector similarity measure for linguistic approximation: Interval type-2 and type-1 fuzzy sets," *Information Sciences*, vol. 178, no. 2, pp. 381–402, 2008.

[2] V. V. Cross and T. A. Sudkamp, *Similarity and compatibility in fuzzy set theory: assessment and applications*, vol. 93. Springer Science & Business Media, 2002.

[3] C.-M. Hwang, M.-S. Yang, W.-L. Hung, and E. S. Lee, "Similarity, inclusion and entropy measures between type-2 fuzzy sets based on the Sugeno integral," *Mathematical and Computer Modelling*, vol. 53, no. 9–10, pp. 1788–1797, 2011.

[4] D. Li, T. Li, and T. Zhao, "A New Clustering Method Based On Type-2 Fuzzy Similarity and Inclusion Measures," *Journal of Computers*, vol. 9, no. 11, pp. 2559–2569, 2014.

[5] L. I. Kuncheva, "Using measures of similarity and inclusion for multiple classifier fusion by decision templates," *Fuzzy sets and systems*, vol. 122, no. 3, pp. 401–407, 2001.

[6] J.-F. Omhover, M. Detyniecki, M. Rifqi, and B. Bouchon-Meunier, "Ranking invariance between fuzzy similarity measures applied to image retrieval," in *2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No. 04CH37542)*, vol. 3, pp. 1367–1372, IEEE, 2004.

[7] B. Bouchon-Meunier, M. Rifqi, and M.-J. Lesot, "Similarities in Fuzzy Data Mining: From a Cognitive View to Real-World Applications," in *Computational Intelligence: Research Frontiers*, vol. 5050 of *Lecture Notes in Computer Science*, pp. 349–367, Springer Berlin Heidelberg, 2008.

[8] M.-Y. Chen and D. A. Linkens, "Rule-base self-generation and simplification for data-driven fuzzy models," *Fuzzy sets and systems*, vol. 142, no. 2, pp. 243–265, 2004.

[9] M. Setnes, R. Babuska, U. Kaymak, and H. R. van Nauta Lemke, "Similarity measures in fuzzy rule base simplification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 28, no. 3, pp. 376–386, 1998.

[10] S.-J. Chen and S.-M. Chen, "Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers," *IEEE Transactions on fuzzy systems*, vol. 11, no. 1, pp. 45–56, 2003.

[11] R. Chutia and M. K. Gogoi, "Fuzzy risk analysis in poultry farming based on a novel similarity measure of fuzzy numbers," *Applied Soft Computing*, vol. 66, pp. 60 – 76, 2018.

[12] J. Navarro, C. Wagner, U. Aickelin, L. Green, and R. Ashford, "Exploring differences in interpretation of words essential in medical expert-patient communication," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2157–2164, IEEE, 2016.

[13] D. Wu and J. M. Mendel, "A comparative study of ranking methods, similarity measures and uncertainty measures for interval type-2 fuzzy sets," *Information Sciences*, vol. 179, no. 8, pp. 1169–1192, 2009.

[14] C. Wagner, S. Miller, and J. Garibaldi, "Similarity based applications for data-driven concept and word models based on type-1 and type-2 fuzzy sets," in *Fuzzy Systems (FUZZ), 2013 IEEE International Conference on*, pp. 1–9, 2013.

[15] M. R. Rajati and J. M. Mendel, "On advanced computing with words using the generalized extension principle for type-1 fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 5, pp. 1245–1261, 2014.

[16] B. Bouchon-Meunier, G. Coletti, M.-J. Lesot, and M. Rifqi, "Towards a conscious choice of a fuzzy similarity measure: a qualitative point of view," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 1–10, Springer, 2010.

[17] C. P. Pappis and N. I. Karacapilidis, "A comparative assessment of measures of similarity of fuzzy values," *Fuzzy Sets and Systems*, vol. 56, no. 2, pp. 171–174, 1993.

[18] S.-M. Chen, M.-S. Yeh, and P.-Y. Hsiao, "A comparison of similarity measures of fuzzy values," *Fuzzy sets and systems*, vol. 72, no. 1, pp. 79–89, 1995.

[19] X. Luo and C. Zhang, "An axiom foundation for uncertain reasonings in rule-based expert systems: Nt-algebra," *Knowledge and Information Systems*, vol. 1, no. 4, pp. 415–433, 1999.

[20] H. B. Mitchell, "On the dengfeng–chuntian similarity measure and its application to pattern recognition," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3101–3104, 2003.

[21] H. Zhang, W. Zhang, and C. Mei, "Entropy of interval-valued fuzzy sets based on distance and its relationship with similarity measure," *Knowledge-Based Systems*, vol. 22, no. 6, pp. 449–454, 2009.

[22] H. Bustince, E. Barrenechea, and M. Pagola, "Relationship between restricted dissimilarity functions, restricted equivalence functions and normal EN-functions: Image thresholding invariant," *Pattern Recognition Letters*, vol. 29, no. 4, pp. 525–536, 2008.

[23] L. Xuecheng, "Entropy, distance measure and similarity measure of fuzzy sets and their relations," *Fuzzy Sets and Systems*, vol. 52, no. 3, pp. 305–318, 1992.

[24] I. Couso, L. Garrido, and L. Sánchez, "Similarity and dissimilarity measures between fuzzy sets: A formal relational study," *Information Sciences*, vol. 229, pp. 122 – 141, 2013.

[25] L. Dengfeng and C. Chuntian, "New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions," *Pattern recognition letters*, vol. 23, no. 1-3, pp. 221–225, 2002.

[26] I. Jenhani, S. Benferhat, and Z. Elouedi, "Possibilistic similarity measures," in *Foundations of Reasoning under Uncertainty*, pp. 99–123, Springer, 2010.

[27] Z. Liang and P. Shi, "Similarity measures on intuitionistic fuzzy sets," *Pattern Recognition Letters*, vol. 24, no. 15, pp. 2687–2693, 2003.

[28] X. Wu, H. Liao, Z. Xu, A. Hafezalkotob, and F. Herrera, "Probabilistic linguistic multimoora: A multicriteria decision making method based on the probabilistic linguistic expectation function and the improved borda rule," *IEEE Transactions on Fuzzy Systems*, vol. 26, pp. 3688–3702, Dec 2018.

[29] W. Zeng and H. Li, "Relationship between similarity measure and entropy of interval valued fuzzy sets," *Fuzzy Sets and Systems*, vol. 157, no. 11, pp. 1477–1484, 2006.

[30] B. Ziółko, D. Emms, and M. Ziółko, "Fuzzy evaluations of image segmentations," *IEEE Transactions on Fuzzy Systems*, vol. 26, pp. 1789–1799, Aug 2018.

[31] A. Tversky, "Features of Similarity," *Psychological Review*, vol. 84, no. 4, pp. 327–352, 1977.

[32] M. De Cock and E. Kerre, "On (un)suitable fuzzy relations to model approximate equality," *Fuzzy Sets and Systems*, vol. 133, no. 2, pp. 137–153, 2003.

[33] F. Klawonn, "Should fuzzy equality and similarity satisfy transitivity? Comments on the paper by M. De Cock and E. Kerre," *Fuzzy Sets and Systems*, vol. 133, no. 2, pp. 175–180, 2003.

[34] B. Bouchon-Meunier, M. Rifqi, and S. Bothorel, "Towards general measures of comparison of objects," *Fuzzy sets and systems*, vol. 84, no. 2, pp. 143–153, 1996.

[35] S. Montes, I. Couso, P. Gil, and C. Bertoluzza, "Divergence measure between fuzzy sets," *International Journal of Approximate Reasoning*, vol. 30, no. 2, pp. 91–105, 2002.

[36] R. Zwick, E. Carlstein, and D. V. Budescu, "Measures of similarity among fuzzy concepts: A comparative analysis," *International Journal of Approximate Reasoning*, vol. 1, no. 2, pp. 221–242, 1987.

[37] B. De Baets, S. Janssens, and H. D. Meyer, "Meta-theorems on inequalities for scalar fuzzy set cardinalities," *Fuzzy Sets and Systems*, vol. 157, no. 11, pp. 1463 – 1476, 2006.

[38] B. De Baets and R. Mesiar, "Metrics and t-equalities," *Journal of mathematical analysis and applications*, vol. 267, no. 2, pp. 531–547, 2002.

[39] D. J. Dubois and H. Prade, *Fuzzy sets and systems: theory and applications*, vol. 144. Academic press, 1980.

[40] M. M. Deza and E. Deza, "Encyclopedia of distances," in *Encyclopedia of Distances*, pp. 1–583, Springer, 2009.

[41] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[42] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons," *Biol. Skr.*, vol. 5, pp. 1–34, 1948.

[43] S.-M. Chen, "A new approach to handling fuzzy decision-making problems," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 18, no. 6, pp. 1012–1016, 1988.